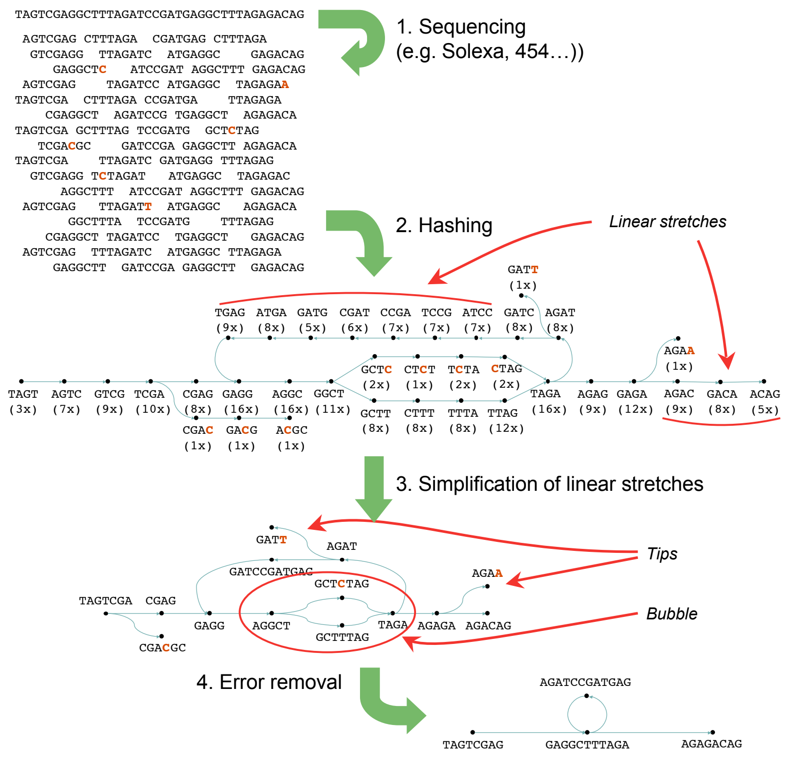
EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

Velvet: *de novo* assembly using very short reads

Daniel Zerbino and Ewan Birney

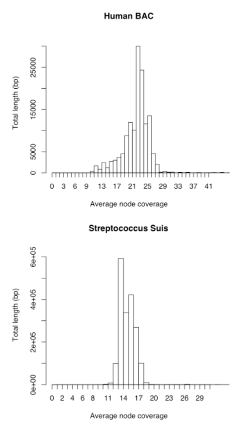


A. Initial pipeline of the Velvet package.

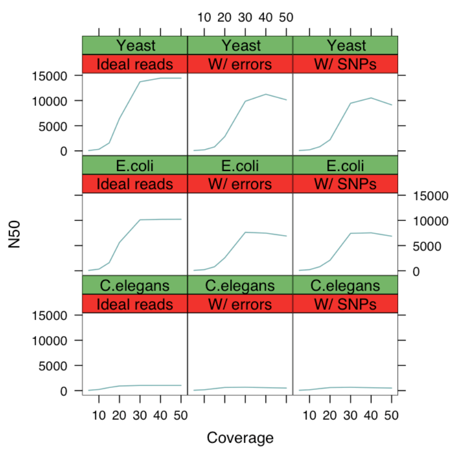
	Human BAC			Streptococcus suis		
Graph:	# nodes	N50 (bp)	Max. length (bp)	# nodes	N50	Max. length (bp)
Initial	309,723	10	10	3,621,167	16	16
Simplified	190,441	10	60	2,222,845	16	44
Tips clipped	5,903	744	4,284	15,267	2,195	7,949
Tour Bus	1,934	1,140	4,624	3,303	4,334	17,811
Cov. cutoff	1,169	1,360	6,568	1,496	8,564	29,856
Ideal	1,120	1,565	7,028	1,305	9,609	29,856

B. Node (~contig) statistics at the different stages of the process described above for the analysis of two experimental datasets generated from a human BAC and Streptococcus suis respectively. The ideal graph is simply obtained by building the graph of the known reference sequence.

C. Coverage distribution of the final contigs produced by the experiments described in B. Analysis of contigs longer than 100 bp showed no misassemblies and error rates of 0.02% and 0.004% respectively.



D. Whole chromosome simulations of sequencing projects on Yeast chr. IV, E.coli, and C. elegans chr. V. The simulations involve respectively error-free reads, reads with errors, and reads with errors applied to a diploid sample.



Background

The Sanger method has long been the dominant approach to sequencing. However, novel technologies, such as 454 and Solexa, are now capable of generating far cheaper, but at the same time far shorter reads (25 to 100 bp instead of 800 to 1000bp). Although greater coverage depths are thus affordable (50-200x instead of 2-10x), de novo genome assembly from those datasets is significantly more complex. Firstly, memory costs are an issue when dealing with so many elements, and secondly, the short read length implies that the assembler must be able to deal with numerous ambiguous overlaps.

We present here an algorithmic package in development, Velvet, which is specifically designed to deal with short read sequencing. Velvet has two tasks: to remove errors from the dataset, and then untangle repeated regions of the genome.

De Bruijn graphs

Given the exceptional redundancy of the information, an appropriate organization of data is crucial. Whereas most assemblers currently use the overlap-consensus-layout approach, where every read is a separate entity, the de Bruijn graph allows us to focus our analysis on observed words (or k-mers). A given k-mer is therefore represented by a unique node, regardless of how many times it was observed.

Furthermore, this representation seamlessly accommodates long and short read mixtures. Long reads are thus represented as paths going from one word to the next, alongside an undifferentiated mass of short reads.

The costliest part of constructing a de Bruijn graph consists in hashing all the reads, according to a given word length. This operation offers nonetheless an advantageous time complexity compared to a general pairwise alignment of all the sequences, especially given the high coverage depths encountered. Once all the reads have been hashed, each of their paths are traced along the k-mer nodes, incrementing coverage and creating the appropriate arcs along the way. Figure A gives the example of an imaginary sequencing project.

Error Removal with Velvet

Velvet's error correction algorithm, Tour Bus, concentrates on removing errors without disrupting connections within the graph. This ensures that a unique point of the genome with low coverage is not arbitrarily destroyed. For this reason, error removal is done after graph creation.

Tour Bus distinguishes two types of errors: tips and bubbles. Tips arise from low quality read ends which do not overlap with anything. Bubbles are created either by an error in the middle of a long read, or by two erroneous read ends accidentally overlapping.

Because they are easily identifiable in the graph, tips are the first to be removed. A tip is considered dubious depending on its length and its coverage.

Bubbles are then detected and removed by a progressive search of the graph similar to Dijkstra's algorithm. This systematic progression through the nodes gives Tour Bus its good time complexity (roughly N*log(N), N being the number of nodes). Bubbles are not corrected by elimination of unwanted data, but by projection of one branch on the other. This means that all the reads and connections belonging to the dubious branch are conserved, and remapped onto the other. Finally, low coverage nodes which have not been assimilated to larger contigs are eliminated.

Experimental Results

Velvet has already proved successful in removing errors and isolating long unique regions from experimental datasets: a human BAC and Streptococcus Suis. Table B and Figure C summarize the results obtained. Velvet's implementation allows it to deal with those datasets in little time (below 10 minutes) and using a reasonable amount of memory (below 16GB)

Whole genome simulations have also been run on C. elegans, E. coli, and yeast. See Figure D for more details.

Future Developments

The major limitation of short reads for de novo assembly is clearly resolving repeats. A number of experimental protocols could however be used to bring in extra information to the assembler:

• Mixing long and short reads. The short reads would therefore eliminate practically all coverage gaps, whereas long reads would help us bridge connections between unique contigs.
• Using short read pairs to extend or even connect unique contigs.
• Pooling fosmids would also allow us to localize our graph construction, thus reducing the effect of distant repeats.

Conclusions

Velvet has shown that handling huge amounts of short reads is computationally feasible, thanks to the use of de Bruijn graphs. The Tour Bus algorithm also takes advantage of this alternative data structure to eliminate errors quickly without sacrificing low coverage areas, thus maintaining the integrity of the graph. Finally, although resolving ambiguous repeats from short read data is a difficult challenge, a number of available options remain to be tested using the Velvet framework.

Acknowledgments

DZ conceived the algorithms in collaboration with EB. The programming and analysis was done by DZ. Both DZ and EB are funded by EMBL central funds. We wish to thank Tony Cox and Zemin Ning for allowing us to use their data. DZ wishes to thank Alison Meynert, Benedict Paten and Michael Hoffman for their ideas and support.

Daniel Zerbino
PhD student
Ensembl
zerbino@ebi.ac.uk

EMBL- EBI
Wellcome Trust Genome Campus
Hinxton
CB10 1SD, Cambridge
United Kingdom

T +44 (0) 1223 494612
F +44 (0) 1223 494468
http://www.ebi.ac.uk